APPENDIX Y:
SCIENCE PERSON-FIT ANALYSIS TECHNICAL REPORT

Scale Comparability for Accommodated Forms in the Rasch Model: A Person-fit Approach

Dong Gi Seo
Michigan Department of Education

Abstract

The purpose of this study is to provide a valid approach and evidence of scale comparability for accommodated form as compared with the non-accommodated forms in a large-scale state assessment by using differential person functioning (DPF) as types of modeling-data misfit (person-fit). Although differential item functioning (DIF) is often investigated to rule out test unfairness as an explanation for the group difference, the unequal sample size and the size of samples have a significant effect on the statistical power of DIF detection procedure. One of the advantages of person-fit analysis is that even with small sample sizes, it can be applied to investigate the scalability of examinees within different subgroups. This study demonstrates that person-fit analysis is appropriate for scale comparability study regardless of the size of the subgroups. Implications and application of the person-fit method to a larger context such as to examine the test equity and comparability for the purpose of program evaluation and peer review are discussed.

Key words: person-fit, differential person functioning, model-fit, test score comparability

INTRODUCTION

Test equity and fairness should be required and ensured to implement standard and assessment system for high stake tests. In order to enhance equity and improve assessment, it is important that the assessment instruments be unbiased. Many studies have relied on large differences in mean score across groups to indicate the presence of bias (e.g., Rosser, 1989). However, Camilli and Shepard (1994) argued that a group mean difference is not sufficient evidence of bias because it may reflect some real group differences. While differential item functioning (DIF) analysis controls for ability when DIF is detecting items that exaggerate the ability difference across groups of examinees.

DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the knowledge in a specific subject area across similar groups of examinees. A large DIF value means that the item is more likely to be measuring additional constructs that function differently from one group to another (Camilli & Shepard, 1994; Roussos & Stout, 1996). The DIF analysis is based on the assumption that test takers who have similar knowledge should perform in similar ways on individual test items regardless of their sex, race, or ethnicity. Many researchers have used DIF analyses as construct comparability in international, comparative, and cross-cultural research (e.g., Kristjansson, Desrochers, & Zumbo, 2003; Hambleton, Merenda, & Spielberger, 2006). These studies have investigated the comparability of translated and/or adapted measures. There are two broad classes of DIF detection methods: Mantel-Haenszel (M-H) (Mantel & Haenszel, 1959) and logistic regression (LogR) approaches. M-H is based on

estimating the probability of a member of the reference or the focal group at a certain ability level getting an item correct. The LogR class of methods (Swaminathan & Rogers, 1990) entails conducting a logistic regression analysis for each item wherein one tests the statistical effect of the grouping variables and the interaction of the grouping variable and the total score after conditioning on the total score. Both M-H and the LogR DIF detection methods have been applied to large scale testing programs.

However, Narayanan and Swaminathan (1996) mentioned that the statistical power of DIF is positively related to the sample size. DIF is not available or appropriate unless: 1) the two forms being compared have the same common items; 2) the two forms have the approximately equal sample size. The unequal sample size and the size of samples have a significant effect on the statistical power of DIF detection procedure (Awuor, 2008). Furthermore, DIF has a trivial impact on the assessment of group mean differences or any particular examinee's score. In a parallel fashion with DIF, differential person functioning (DPF) can be defined as an examination of group differences when person-fit is defined as unexpected differences between the observed and expected performance of persons on a set of items (Engelhard, 2009).

A major advantage of item response theory (IRT) models is the capability of evaluating the fit of a specific IRT model to an examinee's responses (Weiss & Davison, 1981). After IRT models were established in the psychometrics, interests in item-fit and person-fit analysis were intensified in educational and psychological area. If an examinee's item responses show lack of fit, it can be stated as the degree of misfit. Person-fit analysis

under dichotomous IRT models has been developed using several "person-fit" indices. Possible causes for misfitting item responses include a variety of test-taking behaviors (Levine & Rubin, 1979). Low motivation to take the test, guessing, cheating, random responding, deliberate distortion, cultural bias, and misunderstanding the test directions would be examples of the many possible reasons.

Various person-fit statistics and indices have been proposed to detect misfitting examinees (Drasgow & Levine, 1986; Meijer, 1994; Tatsuoka 1984). Several researchers used the person response function as a person-fit index and compared it with other person-fit indices (Trabin & Weiss, 1983; Nering & Meijer, 1998). Meijer and Sijtsma (2001) reviewed a large number of statistics invented for the purpose of identifying non-fitting response pattern. Karabatsos (2003) compared the 36 person-fit indices under different testing conditions to obtain as to which ones are most useful to detect aberrant response. After person-fit analysis was initiated in the early part of the century, person-fit indices have been used to identify unusual test behaviors such as cheating and guessing (e.g., Wright & Stone, 1979). In addition to the application of detecting misfitting responses, person-fit can be applied to cross-cultural studies to investigate the scalability of examinees with different ethnic backgrounds (Reise & Flannery, 1996). Nevertheless, few studies have used the person fit analysis to investigate the scalability of examinees with different group levels. Engelhard (2009) used item-fit and person-fit as a part of broader equivalence framework that would include an examination of the conceptual or functional equivalence of items and scores as well an examination of the equivalence of

construct as operationally defined within each student groups and modification conditions.

Purpose

One of practical advantages of person-fit analysis is that the quite small sample size of the subgroups can be applied to investigate the scalability of examinees within different subgroups. Furthermore, person-fit analysis does not require the common items between two forms given that the same model is applied to the two forms. On this point, the purpose of this study is to guide a more practical basis for the application of person-fit index (i.e., OUTFIT in the context of Rasch Model) to check whether a science test scores from a high-stakes large scale state assessment (Grade 5 and 8) are comparable between the non-accommodated and accommodated forms, and between the non-translated and translated forms. Specifically, the results from this person fit analysis were expected to provide additional construct validity evidence in that positive results are indicative of comparability of scores across accommodated forms assumed to be on the same scale of measurement.

Methods

Data Resources

Data from a high-stakes science assessment for Grade 5 and Grade 8 in northeastern state of the US were used to investigate scale comparability. Table 1 describes the demographic characteristics of the Grade 5 and Grade 8 students.

Insert Table 1

In this science test each form consists of operational and field test items. One operational form and each field test form were constructed for elementary school (Grade 5) and middle school (Grade 8). The test structure for science tests is summarized in Table 2.

Insert Table 2

All these forms were built under the same test specification (blueprint) that mapped the state science curricula and standards. Items are classified according to standard strand. In addition to the three major science subject area (life, earth, and physical science), each form must also include a set number of items that address the process skills of constructing and reflecting. Within each grade, the forms are parallel and test scores were post equated.

Test Forms and Items

This study used only operational items for the science assessments for the 5^{th} and 8^{th} grades. Thus, the total number of items used in Form 1 was 48 items for Grade 5, and 52 items for Grade 8. Consequently, the total number of items for all forms was 144 items (32 base items +16 matrix items × 7 forms) for Grade 5, and 172 items (32 base items +20 matrix items ×7 forms) for Grade 8. This study used only Form 1 that included accommodated forms.

Each test form consisted of base items and field-test items. Base items were those items that are the same across all test forms within each subject and grade and count toward a student's score. Field-test items were those being administered for the first time to gather statistical information about the items. Some of them would also be potentially used for linking to future forms for some administrations, and for generating school level scores if they passed the field test item reviews. These items did not count toward an individual student's score for the current test cycle. Technically, for the 2010 administration, the grade 5 and 8 science assessments were equated to Fall 2009 by mean/mean method (Loyd and Hoover, 1980).

Item Development and Selection

In addition to the content coverage requirements, all operational items were reviewed by both the Bias review Committees (BSCs) and the Content Advisory Committees (CACs). These committees sort the field tested items and identify which items are eligible for inclusion in the operational item pool. There is a separate pool for each subject area and grade level assessed. It is from these pools that items are selected to meet the requirements outlined in the assessment blueprints. Test forms are developed using the approved test items. In addition to overarching content requirements for each test form developed, content experts and psychometricians consider requirements related to subdomains, graphics and other visual representations, passage and content dependent items, and clueing concerns.

The statistical process was also preceded with the Data Review Committees, both the BSC and the CAC post field test. The committees evaluated the field test items using

item statistics from classical measurement theory and item response theory models. From the work of these committees, a pool of items that are eligible to be used in constructing the operational forms was identified.

Because the science assessment was used in making individual decisions about students, it must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). The targeted reliability coefficient was .90 (or higher) for science assessment. Other psychometric properties considered in item selection also included item difficulty (p-value ranged from .25 to .95), item discrimination (point biserial correlations were above .25), and differential item functioning (Mantel-Haenszel).

Accommodated Testing Materials

The following accommodated testing materials were provided for the science test: Braille, Large Print, Oral Administration and Bilingual. As an equivalent form (with a standard test form), the accommodated science test forms differ from the standard form tests only in the provided accommodations.

Braille. All test items are screened for adaptability to Braille. If an item not suitable for Braille is selected for use on a base-test form, an appropriate item would be substituted on the Braille form or the item would be dropped from the Braille form.

Large Print. All test items are screened for adaptability to large print. Text is enlarged to one of four font sizes based on the degree of visual impairment. The font sizes offered reflect the sizes of print being used in current instructional situations.

Oral Administration. Students may have oral administrations using a cassette recording of the scripted test or by having a test administrator read the script aloud.

Bilingual Tests (Translated form). Tests are translated into Spanish and Arabic; the top language groups represented in the state after English. All tests include the English questions followed by the translated questions. Students may have tests interpreted on the day of testing for languages where a printed bilingual version is not available.

IRT Model Fit and Item Parameter Calibration

The Rasch (1961) model was used to derive the scale score system for the science test. The Rasch model is useful for scaling students on single or multiple latent proficiencies based on simple structure. The Rasch model is defined via the following mathematical measurement model:

$$p(u_{ij} \mid \theta_j, b_i) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}}, \quad (1)$$

where b_i is the difficulty estimate for item i, and θ_j is estimate of ability of examinee j. Rasch model assumes that attributes of examinees are independent with each other.

The item difficulty parameters were estimated by WINSTEP software.

To maintain the same performance standards across different administrations, all tests must have comparable difficulty. This comparable difficulty was maintained from administration to administration at the total test level and, as much as possible, at the reporting strand level. A post-equating procedure was applied on the science test. This equating design ensured that the level for any performance standard established by the science test on the original test was maintained on all subsequent test forms.

The following Table 3 is a summary of the fixed item difficulties for the Form 1 of the science test.

Insert Table 3

Table 3 also showed the summary of the item fit statistics. All items have the mean square outfit statistics less than 1.5 criteria. Overall, based on the outfit statistics, all items fit the Rasch model.

Person Score Calibration

The students' abilities were estimated by unconditional maximum likelihood based on the fixed-item parameters in Table 3. The students' abilities and OUTFIT mean-square were computed by the WINSTEP software. Table 4 described the summary of students' estimated abilities for the science test.

Insert Table 4

The Person Fit Index

This study used the Outfit mean-square (Wright & Linacre, 2000) as a person-fit statistics. The Outfit mean-Square was used to compare the degree of misfit between non-accommodated and accommodated forms, and the non-translated and translated forms. Outfit mean-square is known as outlier-sensitive fit statistics (Wright & Linacre, 2000). This is more sensitive to unexpected observations by persons on items that are relatively very easy or very hard for them. Outfit mean-square can be defined as

$$MS_{OUTFIT} = \sum_{i=1}^{N} \frac{\left(X_i - P_i\right)^2}{P_i\left(1-P_i\right)} \Big/ N \qquad (2)$$

Where N is the number of items, X_i is the response of item i, and P_i is the expected response of item i. This is based on the conventional chi-square statistic. Outfit mean-square is the chi-square statistic divided by its degrees of freedom. Consequently its expected value is close to 1.0. Values greater than 1.0 (underfit) indicate non-fitting response patterns in the data. Values less than 1.0 (overfit) indicate that students fit well to an expected model. Although there are various interpretation guidelines, one guideline states that values from .5 to 1.5 are fine while values greater than 2 need to inspect the associated person (Wright & Linacre, 2000). This study used the cutoff value of 1.5 (>1.5) to detect non-fitting response pattern for the science test.

RESULTS

Non-accommodated vs. Accommodated Forms

Figure 1- 2 showed the scatter plot of Outfit mean-square by $\hat{\theta}$ level for non-accommodated and accommodated forms of the science test. The theoretical cut value of 1.5 was applied to detect the non-fitting response patterns for students using the non-accommodated and the accommodated forms. Since the person-fit analysis would consider response patterns within the same structural model, a combined response set was required to compare two forms. Since the number of the students who took the accommodated form test was smaller than the number of students who took the non-accommodated form, it was not legitimate to conduct separate analyses for these two groups. Thus, person-fit analysis was applied to compare two forms whether the test is valid regardless of form types under combined data set. Since the person-fit analysis would consider response patterns within the same structural model, it is not necessary to regard the assumption of equal sample sizes for the two forms. Table 2 provided evidence regarding the empirical parallelism of the forms.

Insert Figure 1 and 2

Table 5 shows the summary of Outfit mean-square statistics for the students who took the non-accommodated and accommodated forms, and the non-translated and translated forms.

Insert Table 5

Table 5 shows the summary of Outfit mean-square statistics for the students who took the non-accommodated form and accommodated form group. In the Grade 5, the mean and SD of Outfit mean-square for non-accommodated form was 1.029 and .247. The Outfit mean-square ranged from .21 to 4.07. The mean and SD of Outfit mean-square for the

accommodated form was 1.106 and .223. The Outfit mean square ranged from .40 to 3.86. The type I error (the ratio of non-fitting responses) of the non-accommodated form was .033, while that for the accommodated form was .049.

In the Grade 8, the mean and SD of Outfit mean-square for non-accommodated form was 1.015 and .225. The Outfit mean-square ranged from .29 to 3.19. The mean and SD of Outfit mean-square for the accommodated form was 1.186 and .266. The Outfit mean square ranged from .59 to 2.71. The type I error (the ratio of non-fitting responses) of the non-accommodated form was .034, while that for the accommodated form was .115.

Non-translated vs. Translated Forms

Figure 3-4 showed the scatter plot of Outfit mean-square by $\hat{\theta}$ for non-translated and translated form in the science test. The theoretical cut value of 1.5 was also applied to detect the non-fitting response patterns for students using the non-translated form and the translated form. Again, since the number of the students who took the translated form test was very small as compared with the number of students who took the non-translated form, person-fit analysis was applied to compare two forms whether the test is valid regardless of form types under combined data set.

Insert Figure 3 and 4

Table 5 shows the summary of Outfit mean-square statistics for the students who took the non-translated form and translated form group. In the Grade 5, the mean and SD of Outfit mean-square for non-translated form was 1.049 and .244. The Outfit mean-square ranged from .21 to 4.07. The mean and SD of Outfit mean-square for the translated form was

1.151 and .202. The Outfit mean square ranged from .85 to 1.86. The type I error (the ratio of non-fitting responses) of the non-translated form was .039, while that for the translated form was .059.

In the Grade 8, the mean and SD of Outfit mean-square for non-translated form was 1.056 and .246. The Outfit mean-square ranged from .29 to 3.19. The mean and SD of Outfit mean-square for the translated form was 1.239 and .252. The Outfit mean square ranged from .75 to 2.14. The type I error (the ratio of non-fitting responses) of the non-translated form was .054, while that for the translated form was .112.

Overall, the differences of Type I errors between non-accommodated and accommodated forms were .016 for the Grade 5, .081 for the Grade 8, and the differences of Type I errors between non-translated form and translated forms were .020 for the Grade 5, .058 for the Grade 8. There were no big differences of Typ1 error rates across forms and grades. In conclusion, the results of the person fit analyses provided no evidence to suggest that the tests function differently across forms or that the persons provide response patterns counterintuitive to the IRT model. Therefore, the meaning of scores along the theta continuum and consequently via the linear translation to scale scores are comparable regardless of the group or form.

Discussions

This study used the person-fit analysis as an alternative method to check whether or not a high-stakes state assessment supports the assumption of measurement invariance/scale comparability by comparing non-accommodated forms with accommodated forms, and

non-translated forms with translated forms for the science content area. Based on the person-fit analysis, there was no evidence to suggest a gross violation of the measurement invariance assumption. The misfit ratios of the accommodated form were similar with the non-accommodated form, and those of the translated form were also similar with the non-translated form in the science test. As a consequence, this study provided additional validity evidence that the inferences made based on $\hat{\theta}$ and any subsequent linear transformations of $\hat{\theta}$ are comparable across forms and accommodations. That is, the meaning of the scores at any point along the underlying ability continuum, as measured by the various forms of the assessments, are comparable and equally valid.

In educational and social science field, more often, additional evidence about scale comparability from a high-stakes assessment in the specific subject area is needed for various purpose (e.g., for program evaluation and for peer review), and the evidence often involves in unbalanced sample size group comparisons (e.g., one subgroups size is extremely small as compared with another). Although DIF analysis has been considered as the most popular method to investigate the comparability of translated and/or accommodated measures, it is affected by not only the difference between two group sample sizes but also the size of the samples. Thus, DIF is sometimes limited to apply into real test assessment because most cases do not have an equal sample size. The person-fit can also be used to identify individual examinees who do not fit with the general group norm due to construct-relevant factors (content, cultural or language bias to certain subgroups or individuals) or construct-irrelevant factors (e.g., cheating, deliberate distorting, answer sheet alignment errors, lack of motivation). In addition, person-fit

analysis can be used to investigate scale comparability between two different groups regardless of the size of the subgroups. Thus, this study was expected to help the test practitioners at large in the application of this method in real testing programs and contribute to the research literature by providing with empirical validity evidence for test score comparability purposes.

However, there are still several unresolved questions we seek to address in future studies and hope the field also investigates. The main issue rests in the numerous assumptions made in order to carry out the study and in fact it was the assumptions made with regard to operational testing and scoring that gave rise to the need to conduct such as study. In an ideal situation, we obtain item parameters from large enough samples that they constitute enough to do separate calibration. Students in low-incidence groups will always present this dilemma and to date there is no strong solution. Bootstrapping techniques are a possibility but they rely heavily on the assumption that small sample used to generate them has some ideal and real data qualities we wish to expand to a larger sample. Aberrant and troublesome response pattern will only be exacerbated when the sample is artificially inflated. So in other words, did this study mimic reality or is our testing environment unrealistic and not representative of best practices. We believe the former. In addition to the nuances that the real testing situation brings to light, there are other considerations as well. More serious attention is needed on the model fit analysis before examining person-fit analysis. If your data as a whole do not fit the model, then the item and person fit indices are invalid for their intended use. The model-data fit should be examined before proceeding to the person-fit analysis because person-fit analysis would be legitimate for person measurement given that data fit well to model at

the item measurement level. However, all items in this study were not exactly fit well to Rasch model. One plausible reason for the observed misfit is the inherent degree of multidimensionality in the assessments and quite frankly any assessment of such a broad number of expectations has trouble supporting unidimensionality arguments. A consequence of multidimensionality is that it is more difficult to obtain assessment results that load heavily on the first principal component. Given more complete control over test design and development, it is possible to construct a more unidimensional test that would likely have better item fit (a smaller proportion of items flagged for misfit).

Future studies may be replicated with a desired condition that data fit well to a certain model at the item measurement. Given that there are numerous other person-fit indicators (Karabatsos, 2003) including *lz* index, and there are also different IRT models (e.g., 3PML and non-parameter IRT model) to address this measurement invariance issue, it is expected that, in the future, the use of person-fit analysis to detect misfitting response patterns should be addressed as a part of validity test that would include examination of the functional equivalence of person scores as well as examination of the items through a variety of person-fit indices and measurement models.

References

Awuor, R. A. (2008). Effect of unequal sample sizes on the power of DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST and Mantel-Haenszel Procedures. Unpublished Doctoral Dissertation.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased items*. Newbury

Park, CA: Sage.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59-67.

Engelhard, Jr. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement. 69,* 585-602.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2006). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16 (4),* 277-298.

Kristjansson, E. A., Desrochers, A., & Zumbo, B. D. (2003). Translating and adapting measurement instruments for cross-cultural research: A guide for practitioners. *Canadian Journal of Nursing Research, 35,* 127–142.

Levine, M, V., & Rubin, D.F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22, 719-748.*

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311-314.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show nonuniform bias. *Applied Psychological Measurement 20(3),* 257-274.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22,* 53-69.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321–334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley, California: University of California Press.

Reise, S.P. & Flannery, W.P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9,* 9-26.

Rosser, P. (1989). Gender and testing. ERIC ED 336457. Educational Resources Information Center.

Roussos, L., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33(2), 215-230.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item characteristic curve models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (p. 83-108). New York:

Academic Press.

Weiss, D.J., & Davison, M. (1981). Test theory and methods. *Annual Review of Psychology, 32,* 629-658.

Wright, B. D., & Stone, M.H. (1979). *Best test design: Rasch measurement.* Chicago: Mesa Press.

Wright, B. D., & Linacre, J.M.(2000). Winsteps: Rasch analysis computer program. Chicago: MESA Press.

Table 1
Demographic Characteristics of Students

Grade 5	Non-Accom (N=16,830)	Accom (N=6,551)			Total (N=23,381)
		Non-trans	trans	total	
Gender					
1.Male	8609	4192	46	4192	12801
2.Female	8221	2359	36	2359	10580
Race					
1. American Indian/Alaskan Native	86	77	0	77	163
2. Black	3512	1074	2	1076	4588
3. Hispanic	863	515	38	553	1416
4. Asian	463	173	9	182	645
5. White	11585	4500	32	4532	16117
6. Multiracial	302	129	1	130	432
7. Missing	19	1	0	1	20

Grade 8	Non-Accom (N=16,969)	Accom (N=5,605)			Total (N=22,574)
		Non-trans	trans	total	
Gender					
1.Male	8629	3518	90	3608	12237
2.Female	8340	1913	84	1997	10337
Race					
1. American Indian/Alaskan Native	107	53	1	54	161
2. Black	3445	945	4	949	4394
3. Hispanic	773	362	49	411	1184
4. Asian	471	76	30	106	577
5. White	11879	3881	90	3971	15850
6. Multiracial	271	110	0	110	381
7. Missing	23	4	0	4	27

Table 2
Test Structure for the Fall 2010 Science Tests (Form 1)

Grade	# of Base Items	# of Matrix Items	# of Total OP Items	# of Field Test Item
5	32	16	48	12
8	32	20	52	12

Table 3
Summary of Item Statistics

Grade	N	B parameters				Mean Square Outfit			
		mean	Std.	min	max	mean	Std.	min	max
05	48	-.119	.782	-1.503	1.802	.99	.12	.77	1.27
08	52	.217	.794	-1.510	2.495	.99	.15	.72	1.37

Table 4
Summary of Students' $\hat{\theta}$ for the Science Test

Grade	Form	Mean	SD	Min	Max
5	Non-Accom	1.609	1.070	-2.51	6.25
	Accom	.765	.831	-1.75	6.25
	Non-Trans	1.392	1.079	-2.51	6.25
	Trans	.592	.827	-1.09	2.85
8	Non-Accom	1.452	.894	-1.26	6.69
	Accom	.631	.649	-2.27	4.28
	Non-Trans	1.251	.911	-1.83	6.69
	Trans	.825	.875	-2.27	3.47

Figure 1

Scatter Plot of *Outfit* Mean-square by $\hat{\theta}$ for 2010 Science Test Grade 5

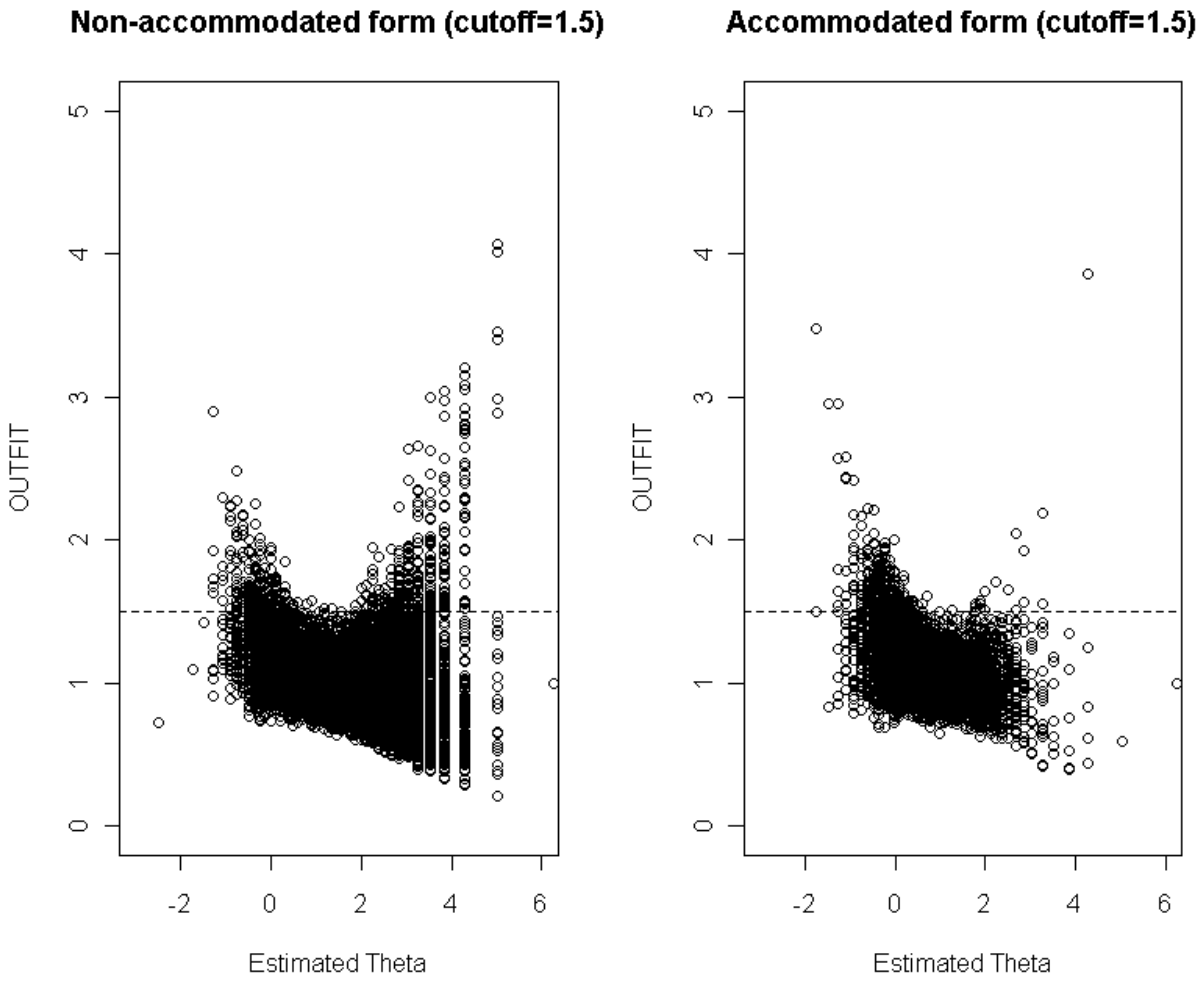

Figure 2

Scatter Plot of *Outfit* Mean-square by $\hat{\theta}$ for 2010 Science Test Grade 8

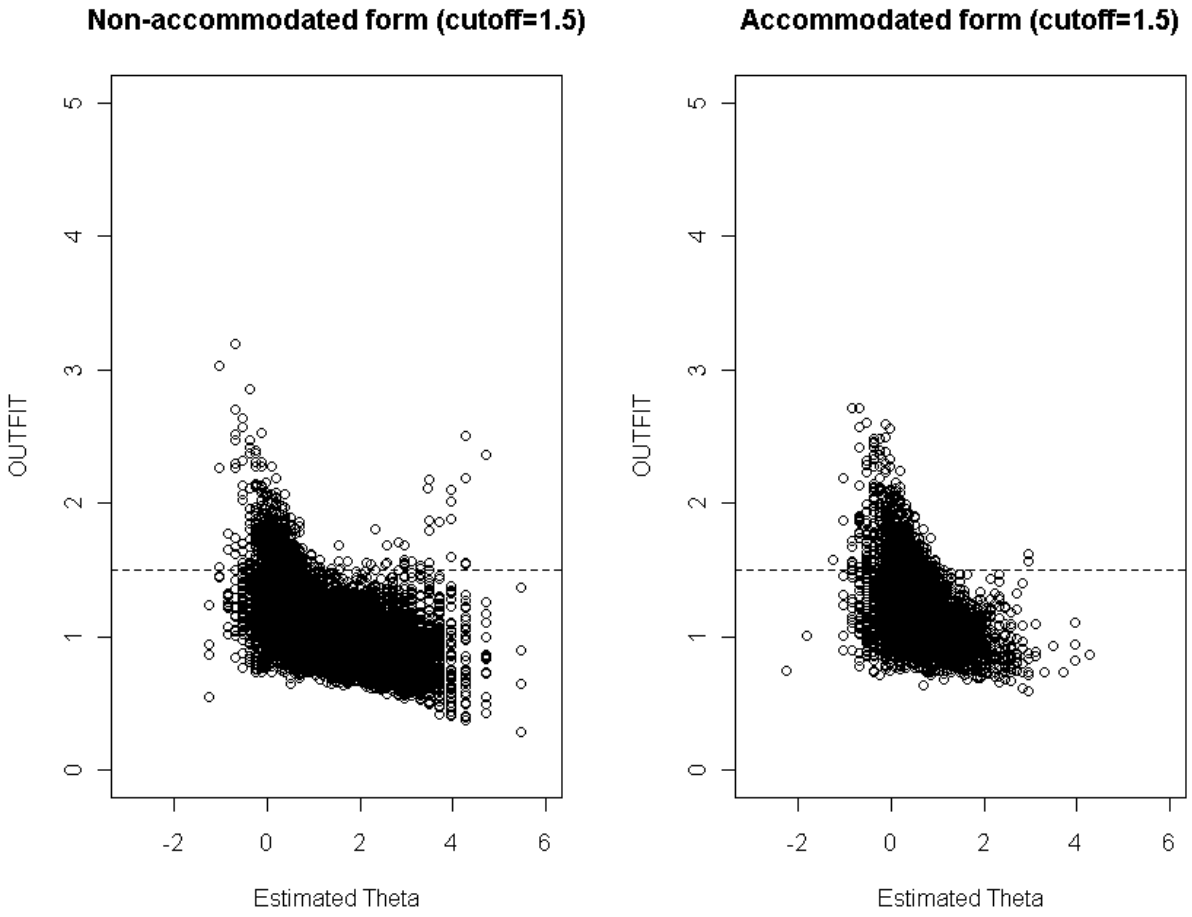

Figure 3
Scatter Plot of *Outfit* Mean-square by $\hat{\theta}$ for 2010 Science Test Grade 5

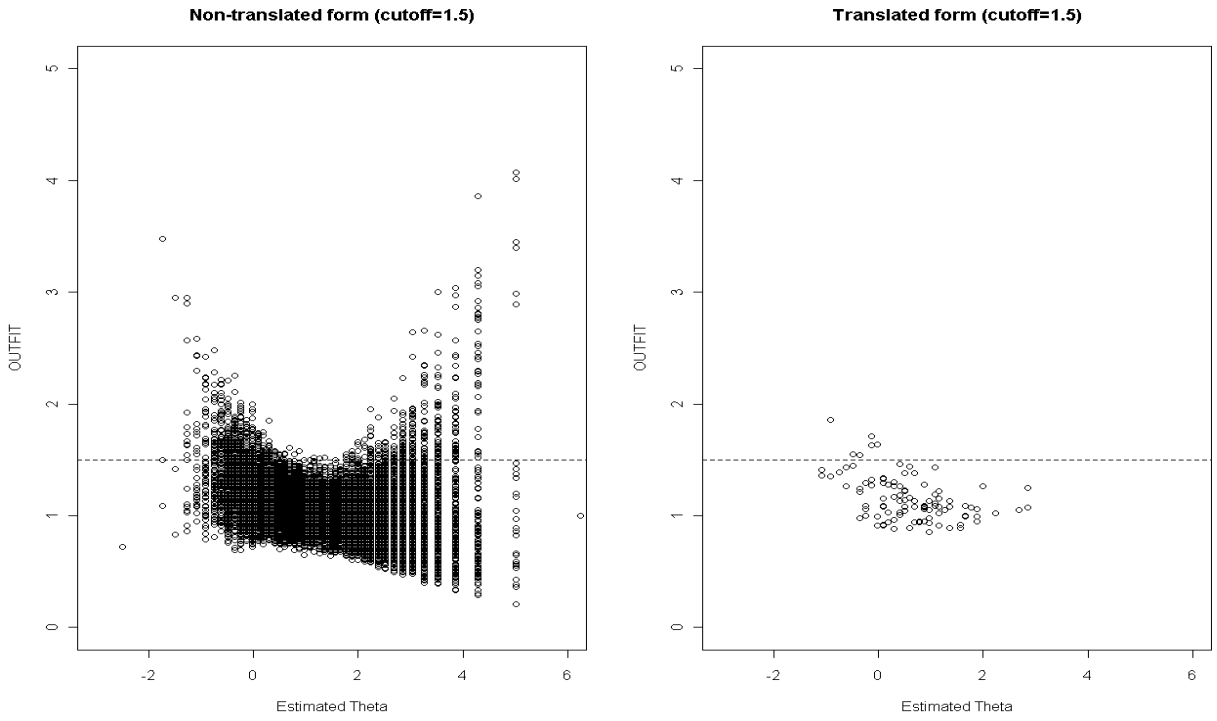

Figure 4
Scatter Plot of Outfit Mean-square by $\hat{\theta}$ for 2010 Science Test Grade 8

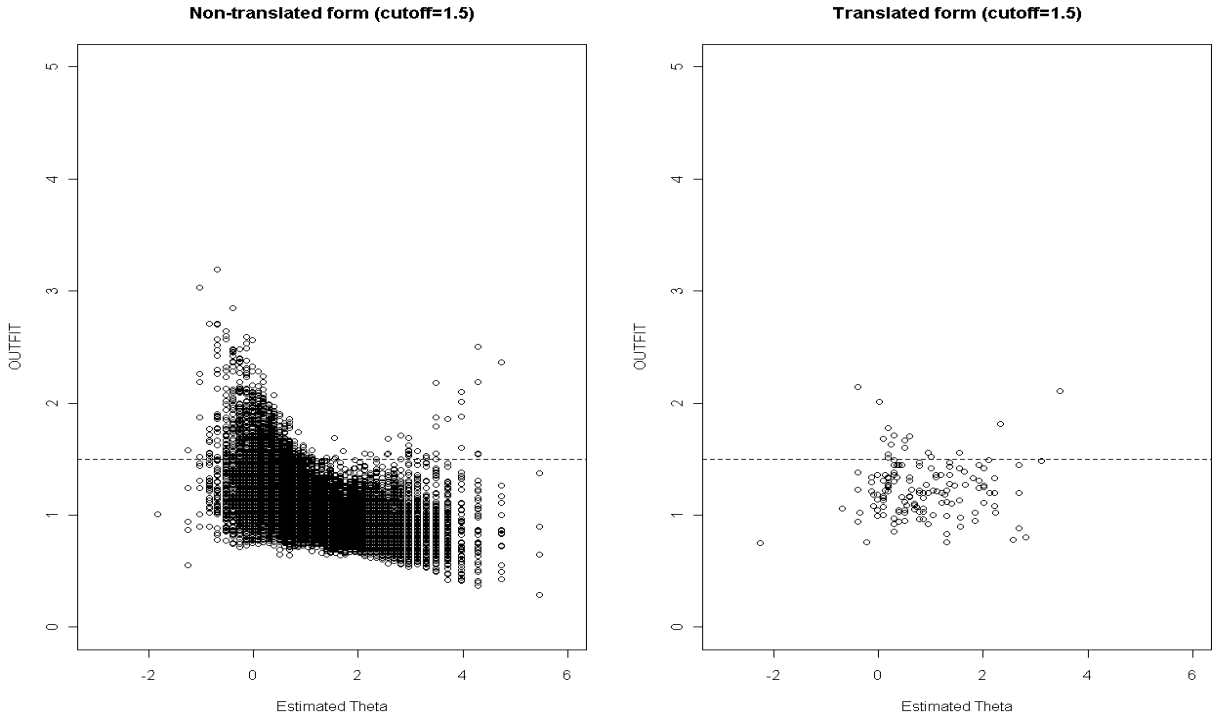

Table 5

Summary of *OUTFIT* mean-square of the Non-accommodated and Accommodated Forms, and Non-translated and Translated forms

Grade	*Form Types*	*Mean*	*SD*	*Min*	*Max*	*Ratio of Misfit*
5	Non-accom	1.029	.247	.21	4.07	.033
	Accom	1.106	.223	.40	3.86	.049
	Non-trans	1.049	.244	.21	4.07	.039
	Trans	1.151	.202	.85	1.86	.059
8	Non-accom	1.015	.225	.29	3.19	.034
	Accom	1.186	.266	.59	2.71	.115
	Non-trans	1.056	.246	.29	3.19	.054
	Trans	1.239	.252	.75	2.14	.112